Das Transkript im Zeitalter der Künstlichen Intelligenz

Automatische Spracherkennung in der Oral History

Tobias Kilgus und Peter Kompiel

1. Ausgangslage

In zahlreichen Archiven, Universitäten, Museen, Gedenkstätten und Bibliotheken liegen bis dato unerschlossene Sammlungen audiovisueller Quellen. Dazu gehören narrativ-biographische Oral-History-Interviews, die „als Musterbeispiel geschichtswissenschaftlicher Forschungsdaten“ (Apel/Leh/Pagenstecher 2022: 208) gelten können. In den Einrichtungen besteht ein großes Interesse, diese wertvollen audiovisuellen Ressourcen technisch aufzubereiten, inhaltlich und multimodal zu erschließen, nach wissenschaftlichen Standards (FAIR) zugänglich zu machen sowie bei neuen („Digital-Born“) Aufzeichnungen die Nachnutzbarkeit von Beginn an zu berücksichtigen.

Eine wichtige Grundlage für die Erschließung von aufgezeichneten Zeitzeug*innengesprächen ist ihre Verschriftlichung in Form von Transkripten. Bislang wurden audiovisuelle Interviews meist manuell transkribiert, was sehr zeitaufwändig und kostenintensiv ist: Sprachlich und technisch versierte Transkripteur*innen müssen die Mediendateien mit einer speziellen Software nach einheitlichen Richtlinien datenschutzkonform und qualitätsgesichert bearbeiten (vgl. Woggon 2012). Der sogenannte „Transkriptionsflaschenhals“ (Brinckmann 2009) hat die Erschließung von Sprachaufzeichnungen aus diesen Gründen lange Zeit verlangsamt.

Eine Möglichkeit zur Verschriftlichung besteht in der Nutzung kommerzieller (automatisierter) Transkriptionsdienste. Oftmals sind diese jedoch mit Blick auf den Daten- und Persönlichkeitsschutz problematisch, da die audiovisuellen Ressourcen in cloudbasierten Umgebungen jenseits lokaler IT-Infrastrukturen verarbeitet werden. Dies bedeutet, dass die Interviewaufzeichnungen, die rechtlich und ethisch sensible audiovisuelle Daten darstellen, sowie die daraus automatisch generierten Transkripte auf externen Servern gespeichert und gegebenenfalls weiterverteilt werden. Häufig fehlt es dabei an Transparenz bei der Nachnutzung der Daten. Zum Beispiel werden die audiovisuellen Ressourcen bei einigen Anbietern ohne das explizite Einverständnis und/oder rechtliche Aufklärung der Nutzer*innen für das Training von Modellen der Künstlichen Intelligenz (KI) eingesetzt. Hinzu kommt, dass die Ergebnisse dieser Dienste, die meist auf andere Use Cases (zum Beispiel die Transkription von Veranstaltungsaufzeichnungen oder Online-Meetings) ausgerichtet sind, häufig nicht den wissenschaftlichen Ansprüchen an Transkriptionsqualität, -richtlinien und -formaten genügen.

Mit der dynamischen Weiterentwicklung der KI in den vergangenen Jahren eröffnen sich vielfältige Möglichkeiten für die Transkription von audiovisuellen Ressourcen mithilfe automatischer Spracherkennung (Automatic Speech Recognition, ASR), die auch in geschichtswissenschaftlichen Kontexten wie der Oral History an Bedeutung

gewinnen. Insbesondere Open-Source-basierte Spracherkennungsanwendungen bergen durch ihre Flexibilität und Community-Unterstützung große Lösungspotenziale für die eingangs skizzierten Herausforderungen.

Der vorliegende Bericht fasst die Erfahrungen des von 4Memory, dem geschichtswissenschaftlichen Konsortium der Nationalen Forschungsdaten-Infrastruktur, im Rahmen der Incubator Funds im Jahr 2024 geförderten Projekts „ASR4Memory“[1] zusammen. An der Universitätsbibliothek der Freien Universität Berlin wurde im Rahmen dieses Projektes ein Prototyp für die automatische Transkription von Oral-History-Interviews und anderen audiovisuellen Forschungsdaten aus der Geschichtswissenschaft entwickelt, der sowohl als Web-Anwendung auf universitätseigenen Servern betrieben wird als auch als Open-Source-Quellcode zur Verfügung steht.

2. Technische Grundlagen

Die Ursprünge der automatischen Spracherkennung reichen bis in die 1950er Jahre zurück. Während ab den 1970er Jahren stochastische Methoden zum Einsatz kamen, setzen sich seit 2010 zunehmend Deep Learning-Verfahren im Bereich der ASR durch.[2] Deep Learning ist ein Teilgebiet des maschinellen Lernens (das wiederum ein Teilgebiet der Künstlichen Intelligenz ist), das auf mehrschichtigen künstlichen neuronalen Netzwerken (KNN) basiert und Modelle mithilfe von Daten trainiert. Bei den Modellen handelt es sich um Algorithmen, die in diesen Daten komplexe Muster erkennen und Vorhersagen oder Entscheidungen auf Grundlage neuer Daten treffen sollen. KNNs können unterschiedlich aufgebaut sein und sich damit in der Art und Weise, wie die Modelle trainiert werden, unterscheiden. Man spricht in diesem Zusammenhang auch von einer „Architektur“. Seit 2017 haben die Transformer-Architekturen in vielen Feldern des Deep Learning an Popularität gewonnen.[3] Die mit Transformern trainierten ASR-Modelle erzielten in den letzten Jahren teils beeindruckende Ergebnisse. Die signifikanten Fortschritte in der ASR sind somit eng mit den jüngsten Entwicklungen im Bereich der Künstlichen Intelligenz verknüpft (vgl. Chollet 2021: 2 ff.; O’Shaughnessy 2024: 14 ff.).

Neben der Entwicklung neuer KNN-Architekturen sind die Verfügbarkeit von leistungsstarker Hardware[4] wie auch die Akkumulation von enormen Datenmengen seit Beginn des Internetzeitalters wichtige Grundpfeiler des KI-Booms (vgl. Chollet 2021: 20 ff.). Für das Training von ASR-Modellen werden klassischerweise Datensätze benötigt, die aus Sprachaufnahmen und den dazugehörigen, meist manuell erstellten

1 Projektwebseite „ASR4Memory”: https://www.fu-berlin.de/asr4memory (25.9.2025).

2 Für eine ausführliche Nachzeichnung der Entwicklung der ASR bis 2005 vgl. Juang/Rabiner 2005.

3 Die Besonderheit hinter der Transformer-Architektur ist der Aufmerksamkeitsmechanismus, der eine parallele Verarbeitung von Daten ermöglicht. Andere gängige Architekturen, wie die sequenziell arbeitenden rekurrenten neuronalen Netze (RNNs), haben Probleme, Kontextinformationen über einen längeren Zeitraum zu speichern, was bei der Verarbeitung von langen Text- bzw. Sprachsequenzen (wie es bei der ASR oftmals der Fall ist) zu Problemen führen kann (vgl. Vaswani et al. 2017: 1 ff.).

4 Für Deep Learning sind vor allem Grafikprozessoren relevant (graphics processing unit, GPU), die im Gegensatz zu klassischen Computerprozessoren (central processing unit, CPU) große Datenmengen parallel verarbeiten können.